

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-148783

(43)Date of publication of application : 30.05.2000

(51)Int.Cl. G06F 17/30

G06F 13/00

(21)Application number : 10-324249

(71)Applicant : NEC CORP

(22)Date of filing : 13.11.1998

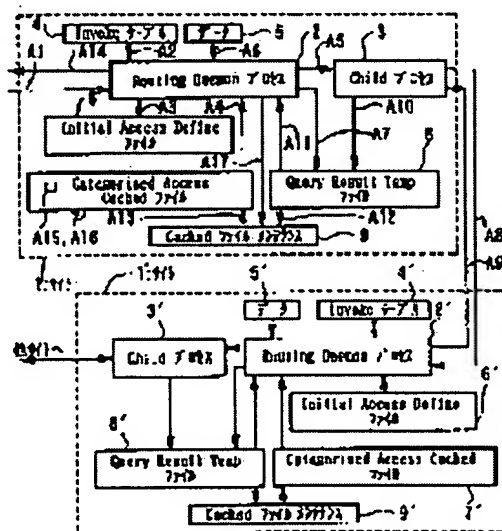
(72)Inventor : KIKUCHI SHINJI

(54) DATA RETRIEVAL DEVICE

(57)Abstract:

PROBLEM TO BE SOLVED: To widen the area of data retrieval using the internet and to make the retrieval efficient.

SOLUTION: When a site 1 receives a data retrieval request A1 from a user, a Routing Daemon process 2 actuates a Child process 3 by referring to data 5 in its device and makes another side 1' outputs a data retrieval request A8. This data retrieval request A8 includes address information A3 on a site to which the data retrieval request is always transferred and address information A4 on a site to which the most likelihood data retrieval request should be transferred from an Initial Access Define file 6. Transfer data A9 retrieved at the other side 1' are held in a Query Result Temp file 8 together with corresponding data A7 by the data 5, merged, and sent as transfer data A14 back to the user. Cached file maintenance 9 extracts the address information on the most likelihood site according to data contents transferred from the Query Result Temp file 8 and updates the Query Result Temp file 8.



(2)

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2000-148783

(P2000-148783A)

(43) 公開日 平成12年5月30日 (2000.5.30)

(51) Int.Cl. ⁷	識別記号	F I	テーマコード(参考)
G 0 6 F 17/30		G 0 6 F 15/40	3 1 0 F 5 B 0 7 5
13/00	3 5 4	13/00	3 5 4 D 5 B 0 8 9
		15/40	3 1 0 C

審査請求 有 請求項の数11 O L (全 10 頁)

(21) 出願番号 特願平10-324249

(22) 出願日 平成10年11月13日 (1998. 11. 13)

(71) 出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72) 発明者 菊地 伸治

東京都港区芝五丁目7番1号 日本電気株式会社内

(74) 代理人 100108578

弁理士 高橋 昭男 (外3名)

Fターム(参考) 5B075 KK02 KK13 KK33 PQ05 PQ20

5B089 GA11 GB09 HA10 JA12 KC15

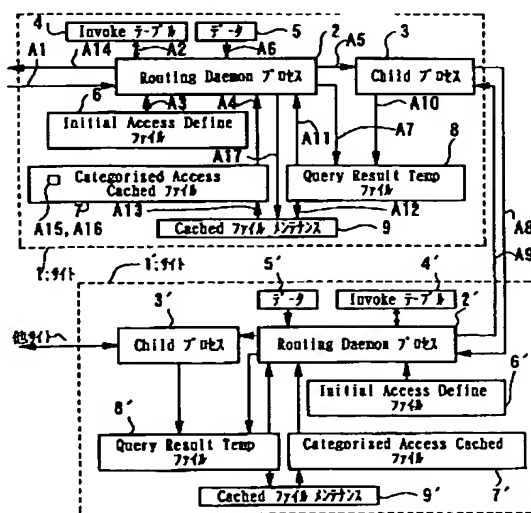
KC23 KC39 KC44

(54) 【発明の名称】 データ検索装置

(57) 【要約】

【課題】 インターネットを利用したデータ検索の広域化、高効率化を図る。

【解決手段】 サイト1が利用者からデータ検索要求A1を受信すると、RoutingDaemonプロセス2は、自装置内のデータ5を参照し、Childプロセス3を起動し、他のサイト1'にもデータ検索要求A8を出力させる。このデータ検索要求A8には、InitialAccessDefineファイル6からの、常にデータ検索要求を転送する サイトのアドレス情報A3と、最も確からしいデータ検索要求を転送すべきサイトのアドレス情報A4を含む。他のサイト1' で検索された転送データA9は、データ5による該当データA7と共にQueryResultTempファイル8に保持され、マージされ、転送データA14として利用者に返される。Cachedファイルメンテナンス9は、QueryResultTemp ファイル8から転送されて きたデータ内容に応じて最も確からしいサイトのアドレス情報を引き出し、QueryResultTemp ファイル8を更新する。



A1: データ検索要求
A2: 情報
A3: IPアドレス情報
A4: ドメイン情報
A5: 引数
A6: サブセット
A7: 該当データ
A8: データ検索要求
A9: 転送データ
A10: 該当データ
A11: 該当データ
A12: IPアドレス
A13: 更新元データ
A14: 転送データ
A15: 相対網番度値
A16: 相対網番度値
A17: 起動要求

【特許請求の範囲】

【請求項 1】 インターネットを利用したデータ検索装置において、複数のデータ転送装置から成り、各データ転送装置は、

データ検索要求を受信すると、自装置内に保持しているデータを参照するとともに、所定数に達するまで、他のデータ転送装置へ前記データ検索要求を転送して当該他のデータ転送装置からデータを受信し、前記自装置内に保持しているデータとマージしてデータ検索要求元に転送することを特徴とするデータ検索装置。

【請求項 2】 インターネットを利用したデータ検索装置において、複数のデータ転送装置から成り、各データ転送装置は、

データ検索要求を受信すると、自装置内に保持しているデータを参照するとともに、所定数に達するまで、他のデータ転送装置へ前記データ検索要求を転送して当該他のデータ転送装置から検索データを受信する常駐の Routing Daemon プロセスと、
該 Routing Daemon プロセスから起動され、複数の他のデータ転送装置とデータ転送に関する通信を行なう Child プロセスとを有することを特徴とするデータ検索装置。

【請求項 3】 前記各データ転送装置は、新たなデータ検索要求を受けると、前記 Routing Daemon プロセスによって起動識別子が登録される Invoke テーブルを備え、Routing Daemon プロセスは、データ検索要求を受けた場合に、該 Invoke テーブルを参照することにより、重複した処理を回避することを特徴とする請求項 2 記載のデータ検索装置。

【請求項 4】 前記各データ転送装置は、前記 Routing Daemon プロセスが、常に前記データ検索要求を転送する他のデータ転送装置のアドレスを格納した Initial Access Define ファイルを備えたことを特徴とする請求項 2 または請求項 3 記載のデータ検索装置。

【請求項 5】 前記各データ転送装置は、前記 Routing Daemon プロセスが、前記データ検索要求のデータ内容に応じて最も確からしいデータ検索要求を転送すべきデータ転送装置のアドレスを格納した Categorized Access Cached ファイルを備えたことを特徴とする請求項 2 ないし請求項 4 のいずれかに記載のデータ検索装置。

【請求項 6】 前記各データ転送装置は、前記データ検索の結果、前記自装置内で参照されたデータおよび他のデータ転送装置から転送された検索データを一時的に保持する Query Result Temp ファイルと、該 Query Result Temp ファイルから転送されてきたデータ内容に応じて最も確からしいデータ転送先のデータ転送装置のアドレスを引き出し、この結果によって前記 Categorized Access Cached ファイルを更新する Cached ファイル メンテナンスとを備えたことを特徴とする請求項 5 記載のデータ検索装置。

【請求項 7】 前記最も確からしいデータ転送装置は、まず、検索条件の指定カテゴリに対し相対網羅度が最小のものを選択し、次に、相対新規度が最小のものを第 1 優先条件、前記相対網羅度が最小のものを第 2 優先条件としてソートして決定することを特徴とする請求項 5 または請求項 6 記載のデータ検索装置。

【請求項 8】 前記最も確からしいデータ転送装置の決定方法は、計算上、扱うべきデータ転送装置の数を制限するサイト上限値設定手順と、保持している検索データ件数が多い順にソートして前記制限されたデータ転送装置の数に相当する順位迄のデータ転送装置を選択する該当サイト選択手順と、検索結果の対象をソートし、同じ対象に関する記述を持つ異なるデータ転送装置の数を前記対象の各々に対して数える該当オブジェクト選択手順と、前記選択された全データ転送装置に対して、各々、データ転送装置の数を変数 X とし、自身を含めた X 台のデータ転送装置に含まれる前記対象の度数に関するヒストグラムを作成するヒストグラム作成手順と、該作成されたヒストグラムを基に網羅度および新規度を計算する網羅度・新規度算出手順と、前記網羅度の小さい順にソートしてデータ転送装置を並び替え、小さい順に前記相対網羅度を付与する相対網羅度付与手順と、前記新規度の小さい順にソートしてデータ転送装置を並べ替え、小さい順に前記相対度を付与する相対新規度付与手順と、該付与された相対網羅度、相対新規度と、前記 Categorized Access Cached ファイルで管理されている現在の相対網羅度、相対新規度との相対平均をとり、前記 Categorized Access Cached ファイルに書き込む更新手順とを含むことを特徴とする請求項 7 記載のデータ検索装置。

【請求項 9】 前記 Child プロセスは、起動の際、データ検索要求を転送すべき他のデータ転送装置のアドレス情報および他のデータ転送装置から転送されてきたデータを一時的に保持すべき前記 Query Result Temp ファイルの名称情報を前記 Routing Daemon プロセスから引数として受け取ることが特徴とする請求項 2 ないし請求項 8 のいずれかに記載のデータ検索装置。

【請求項 10】 前記データ検索要求は、転送される際に、通過するデータ転送装置の Name 特定情報および累積 Hop 数が内部にシーケンス状に記録されることを特徴とする請求項 1 ないし請求項 9 のいずれかに記載のデータ検索装置。

【請求項 11】 前記データ検索要求、Invoke テーブル、Initial Access Define ファイル、Categorized Access Cached ファイルおよび Query Result Temp ファイル BNF (Backus Normal Form) で定義される記述形式で表されることを特徴とする請求項 1 ないし請求項 10 のいずれかに記載のデータ検索装置。

【発明の詳細な説明】**【0001】**

【発明の属する技術分野】 本発明は、インターネット上

のデータ検索要求に対するデータ転送装置に関する。

【0002】

【従来の技術】従来のこの種の転送方式を図4に示す。この方式は、米国論文 (M. A. sheldon, A. Duda, R. Weiss, D. K. Gifford: Discover: A Resource Discovery System based on Content Routing, The 3rd International world-wide web Conference April 10-14, 1995 (<http://www.igd.de/www/www95/papers/82/html-files/discover.html>) (1995).) に記されたものである。

【0003】この従来方式は、利用者に対してサーバの役割を果たす Httpd プロセス 11, ルーティング機能を提供する複数のコンテンツルータ 12 等, 各コンテンツルータが参照し、ルーティングに関する情報を管理する複数のルーティングデータベース 13 等, 各コンテンツルータが参照し、利用者に正しい質問入力を促進するための複数の Refinement データベース 14 等, 検索機能を提供する複数のサーチモジュール 16, ドキュメントデータを管理するためのシステムである複数の WAIS プロセス 17 および各 WAIS プロセス 17 が参照する複数の WAIS データベース 18 を含んでいる。

【0004】利用者が Web ブラウザ等を介して、Httpd プロセス 11 にドキュメントの検索要求 B1 を発行すると、Httpd プロセス 11 は、ルーティング機能を提供し、直接関連しているコンテンツルータ 12 に検索要求 B12 を発行する。このコンテンツルータ 12 は、通常、Httpd プロセス 11 のバックプロセスとして機能する。

【0005】その後、コンテンツルータ 12 は、検索要求 B1 で指定され、検索要求 B2 に含まれる質問語彙妥当か否かを判定するため、Refinement データベース 14 上の正しい質問データ B3 を参照し、妥当な表現に修正する。

【0006】次に、コンテンツルータ 12 は、ルーティングデータベース 13 に問い合わせ、質問語彙に対応して、該当する検索結果を応答できる、もしくはその仲立ちができる他のコンテンツルータ 12', 12'' の位置情報 B4 を得る。この情報はルーティングデータベース 13 の内部ではコンテンツラベル 15 として管理されている。また他のルーティングデータベース 13', 13'' でも同様にコンテンツラベル 15', 15'' が存在している。

【0007】次に、コンテンツルータ 12 は、コンテンツラベル 15 から得た位置情報 B4 が指し示す他のコンテンツルータ 12', 12'' に、検索要求 B2 と等価な検索要求 B5, B5' を転送する。

【0008】他のコンテンツルータ 12', 12'' においても、コンテンツルータ 12 と全く同様の処理を行なう。その際、もしこのコンテンツルータ 12', 12'' がルーティングの際に、最終位置にある場合、直接関連するサーチモジュール 16, 16' に検索要求 B5, B

5' と等価な検索要求 B6, B6' を転送する。サーチモジュール 16, 16' が検索要求 B6, B6' を受けると、この検索要求 B6, B6' と等価な検索要求 B7, B7' を配下のドキュメントデータを管理する WAIS プロセス 17, 17' に発行する。それぞれの WAIS プロセス 17, 17' は検索要求 B1 で指定された質問語彙を含む B9, B9' を各 WAIS データベース 18, 18' に発行し、関連するドキュメントの位置情報 B10, B10' を WAIS プロセス 17, 17' に返す。WAIS プロセス 17, 17' は、対応するサーチモジュール 16, 16' に検索要求 B7, B7' の応答として、ドキュメントの位置情報 B10, B10' と等価な位置情報 B11, B11' を返す。

【0009】同様にサーチモジュール 16, 16' は対応するコンテンツルータ 12', 12'' に検索要求 B6, B6' の応答として、ドキュメントの位置情報 B11, B11' と等価な位置情報 B12, B12' を返す。その後コンテンツルータ 12', 12'' は、起点となるコンテンツルータ 12 に検索要求 B5, B5'' の応答として、ドキュメントの位置情報 B12, B12' と等価な位置情報 B13, B13' を返す。

【0010】起点となるコンテンツルータ 12 は、配下のコンテンツルータ 12', 12'' から検索結果であるドキュメントの位置情報 B13, B13' を全て受け取ると、それらを合成し最終的な位置情報 B14 として Httpd プロセス 11 に戻す。その後、利用者が発行したドキュメントの検索要求 B1 に対応したドキュメントの位置情報 B15 が利用者の Web ブラウザ等上に表示されることになる。

【0011】コンテンツラベル 15, 15', 15'' を更新する場合は、コンテンツルータ 12, 12', 12'' は検索条件を限定しない検索要求 B16, B16' を発行する。その後、前述の手順に応じて、WAIS データベース 18, 18' 上に管理された全ドキュメントの位置情報 B17, B17' が返されることになる。全ドキュメントの位置情報 B17, B17' の内容はコンテンツラベル 15, 15', 15'' を更新し得るだけの内容を持ち、情報内容を持って、コンテンツルータ 12, 12', 12'' はコンテンツラベル 15, 15', 15'' を更新する。

【0012】

【発明が解決しようとする課題】上述した従来の方式 2'' では、図4に示した様に WAIS プロセス 17, 17' を使っているため、ドキュメント検索の際、語彙検索が中心となる。また、基本的に WAIS データベース 18, 18' で管理されるコンテンツデータを対象としており、一般的なレガシーデータを対象とはしていないという第1の問題点がある。

【0013】また、コンテンツラベル 15, 15', 15'' を更新する手段を用意してはいるが、検索の際には

最適化した検索が行われる機構が特になく、決められたコンテンツルータ12', 12" にのみ、常に検索要求をルーティングすることしか出来ないという第2の問題点がある。

【0014】また、仮に、2つのルーティングデータベース13', 13" 内の各々のコンテンツラベル15', 15" 上に、互いに相手の位置情報が登録されている場合には、検索が2重に実施されることになり検索上のロスが発生するという第3の問題点がある。

【0015】従って、本発明の目的は、近年、益々、重要になって来ているインターネット上のドキュメントを含むデータベースの検索を出来るだけ広域にかつ無駄なく、実施出来るデータ転送装置を提供することにある。

【0016】また、本発明の他の目的は、利用者の欲している情報をできるだけ定量的に評価して効率を高めることができるデータ検索装置を提供することにある。

【0017】

【課題を解決するための手段】そこで、本発明の第1のデータ検索装置は、インターネットを利用したデータ検索装置において、複数のデータ転送装置から成り、各データ転送装置は、データ検索要求を受信すると、自装置内に保持しているデータを参照するとともに、所定数に達するまで、他のデータ転送装置へ前記データ検索要求を転送して当該他のデータ転送装置からデータを受取り、前記自装置内に保持しているデータとマージしてデータ検索要求元に転送することを特徴とする。また、本発明の第2のデータ検索装置は、インターネットを利用したデータ検索装置において、複数のデータ転送装置から成り、各データ転送装置は、データ検索要求を受信すると、自装置内に保持しているデータを参照するとともに、所定数に達するまで、他のデータ転送装置へ前記データ検索要求を転送して当該他のデータ転送装置から検索データを受理する常駐のRouting Daemon プロセスと、該Routing Daemon プロセスから起動され、複数の他のデータ転送装置とデータ転送に関する通信を行なうChildプロセスとを有することを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記各データ転送装置は、新たなデータ検索要求を受けると、前記Routing Daemon プロセスによって起動識別子が登録されるInvoke テーブルを備え、Routing Daemon プロセスは、データ検索要求を受けた場合に、該Invoke テーブルを参照することにより、重複した処理を回避することを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記各データ転送装置は、前記Routing Daemon プロセスが、常に前記データ検索要求を転送する他のデータ転送装置のアドレスを格納したInitial Access Define ファイルを備えたことを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記各データ転送装置は、前記Routing Daemon プロセスが、前記データ検索要求のデータ内容に

応じて最も確からしいデータ検索要求を転送すべきデータ転送装置のアドレスを格納したCategorized Access Cached ファイルを備えたことを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記各データ転送装置は、前記データ検索の結果、前記自装置内で参照されたデータおよび他のデータ転送装置から転送された検索データを一時的に保持するQuery Result Temp ファイルと、該Query Result Temp ファイルから転送されてきたデータ内容に応じて最も確からしいデータ転送先のデータ転送装置のアドレスを引き出し、この結果によって前記Categorized Access Cached ファイルを更新するCached ファイル メンテナンスとを備えたことを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記最も確からしいデータ転送装置は、まず、検索条件の指定カテゴリに対し相対網羅度が最小のものを選択し、次に、相対新規度が最小のものを第1優先条件、前記相対網羅度が最小のものを第2優先条件としてソートして決定することを特徴とする。さらに、本発明のデータ検索装置の好ましい実施の形態は、前記最も確からしいデータ転送装置の決定方法は、計算上、扱うべきデータ転送装置の数を制限するサイト上限値設定手順と、保持している検索データ件数が多い順にソートして前記制限されたデータ転送装置の数に相当する順位迄のデータ転送装置を選択する該当サイト選択手順と、検索結果の対象をソートし、同じ対象に関する記述を持つ異なるデータ転送装置の数を前記対象の各々に対して数える該当オブジェクト選択手順と、前記選択された全データ転送装置に対して、各々、データ転送装置の数を変数Xとし、自身を含めたX台のデータ転送装置に含まれる前記対象の度数に関するヒストグラムを作成するヒストグラム作成手順と、該作成されたヒストグラムを基に網羅度および新規度を計算する網羅度・新規度算出手順と、前記網羅度の小さい順にソートしてデータ転送装置を並べ替え、小さい順に前記相対網羅度を付与する相対網羅度付与手順と、前記新規度の小さい順にソートしてデータ転送装置を並べ替え、小さい順に前記相対新規度を付与する相対新規度付与手順と、該付与された相対網羅度、相対新規度と、前記Categorized Access Cached ファイルで管理されている現在の相対網羅度、相対新規度との相対平均をとり、前記Categorized Access Cached ファイルに書き込む更新手順とを含むことを特徴とする。

【0018】

【発明の実施の形態】次に、本発明の実施の形態について説明する。図1は、本発明のデータ内容に応じたルーティング方式を採用したデータ転送装置の実施例を示す。本実施例では、データ転送装置をサイト (Site) と称し、図1には2つのサイト1と1' が示されている。

【0019】サイト1にはデータ検索要求を受けると、自装置内に保持しているデータを参照すると共に他サイ

ト 1' へ同一のデータ検索要求を転送・発行し、その結果、他サイト 1' 内に保持しているデータを受理する常駐プログラムの Routing Daemon プロセス 2 と、Routing Daemon プロセス 2 から起動され、他の、サイト 1' 等の複数のデータ転送装置と実際のデータ転送に関する通信を行なう Child プロセス 3 と、データ検索要求の処理状況を管理・記録する Invoke テーブル 4 と、サイト内で固有に保持されているデータ 5 と、Routing Daemon プロセス 2 が常にデータ検索要求を転送・発行するサイトのアドレスを記した Initial Access Define ファイル 6 と、後述の該当データマイニング方式により更新され、データ検索要求を転送・発行する際に最も確からしいサイトのアドレスを得るのに使用する Categorized Access Cached ファイル 7 と、データ検索要求によるデータ 5 への参照結果であるデータ、並びに上述のデータ検索要求の結果、転送されて来たサイト 1' 内に保持されているデータ 5' を一時的に保持・記録する Query Result Temp ファイル 8 と、データマイニング方式により、Query Result Temp ファイル 8 からデータ検索要求の条件に応じて最も確からしいデータ転送先のサイトのアドレスを引き出し、その結果をもって Categorized Access Cached ファイル 7 を更新・維持管理する Cached ファイ

ル メンテナンス 9 を含んでいる。

【0020】Invoke テーブル 4 は、各サイトに 1 つ設定されるものであり、サイト間でやりとりされるデータ検索要求の処理状況を管理・記録するものである。Routing Daemon プロセス 2 がデータ検索要求 A_1 を受理すると、Invoke テーブル 4 をアクセスし、該当する起動識別子（以下 InvokeID）を含んだ情報 A_2 が存在するか否かを確認する。この InvokeID を含んだ情報 A_2 を初めて扱う場合は、InvokeID を含んだ情報 A_2 をテーブル 4 に書き込むことで登録する。それに対して、サイト 1 がサイト 1' 内の Child 3' から、既に同一の InvokeID に相当するデータ検索要求 A_1 を受けている場合は、その InvokeID を含んだ情報は、登録済みになっているので、新たに同じデータ検索要求を受けた場合、Routing Daemon プロセス 2 が処理を無視する、もしくは拒否する等を行ない、2 重に同一の処理が実施されることを防止する。

【0021】Invoke テーブル 4 は、以下の BNF (Backus Normal Form) で定義される記述形式を持ち、同一のデータ検索要求は容易に判定出来る。

【0022】

【数 1】

```

InvokeTable ::= <originl_pass> <invokeID>;
<originl_pass> ::= <identifier>;
<identifier> ::= <ip_string> | <dns_string> | <other>;
<ip_string> ::= {<number>3} { '.' <number>3}3;
<dns_string> ::= {<word> '.'} {<word> '.'} *;
<other> ::= <word> *;
<string> ::= <character> | <character> <character> *;
<word> ::= <alphabetic> {<alphabetic> *} ;
<character> ::= <alphabetic> | <number> | <special>;
<number> ::= 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0;
<special> ::= < | > | = | + | - | * | / | & | ^ | ~ | _ | @ | $ | % | : | . | ! | ? ;
<invokeID> ::= <number>15;

```

【0023】また、データ検索要求 A_1 は、以下の BNF で定義される記述形式を持ち、転送の際に通過サイトの Name 特定情報、並びに累積 Hop 数が内部に記憶される。尚、通過サイトの Name 特定情報は、順序が把握出

来るようにシーケンス状に記録される。

【0024】

【数 2】

```

QueryText ::= <originl_pass> <invokeID> <query> <hop_number> <pass_list> <local_pass>;
<hop_number> ::= <number> 3;
<pass_list> ::= ' | ' { <remote_pass> ' | ' } *;
<remote_pass> ::= <identifier>;
<local_pass> ::= <identifier>;
<originl_pass> ::= <identifier>;
<identifier> ::= <ip_string> | <dns_string> | <other>;
<ip_string> ::= { <number> 3 } { ' . ' <number> 3 } 3;
<dns_string> ::= { <word> ' . ' } { <word> ' . ' } *;
<other> ::= - <word> *;
<query> ::= <string>;
<string> ::= <character> | <character> <character> *;
<word> ::= <alphabetic> { <alphabetic> } *;
<character> ::= <alphabetic> | <number> <special>;
<number> ::= 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0;
<special> ::= ( | ) | = | + | - | * | / | & | ` | ~ | _ | @ | $ | % | : | . | ! | ?;
<invokeID> ::= <number> 15;

```

【0025】Routing Daemon プロセス2は、その後、categorized Access Cached ファイル7をアクセスし、先のデータ検索要求A₁を転送・発行するべきサイトのアドレス情報A₄を得ると共に、Initial Access Define ファイル6もアクセスし、同様に先のデータ検索要求A₁を転送・発行するサイトのアドレス情報A₃を得る。Initial Access Define ファイル6には、Routing Daemon プロセス2が常にデータ検索要求A₁を転送・発行するサイトのアドレスが記されているのに対して、Categorized Access Cached ファイル7には、データ検索要求A₁で指定された検索条件としてのカテゴリに応じて、データ検索要求を転送・発行する際に最も確からしいサイトのアドレスが記されている。

【0026】その際の最も確からしいサイトの決定手順は、図2に記されたようになる。ステップS1として、まず、後述の相対網羅度A₁₅が最小のものを、利用者が検索条件として指定したカテゴリに対し、最も有効と思われるサイトとして選択する。相対網羅度A₁₅が最小ということは、もっとも多くの該当するデータを保持しているサイトを意味する。

【0027】次に、ステップS2では、利用者が検索条件として、指定したカテゴリに対し、以下の優先条件でソートし、複数のサイトを規定数分、選択する。

第一優先条件) 相対新規度A₁₆が最小のもの

第二優先条件) 相対網羅度A₁₅が最小のもの

この条件設定の理由は、可能な限り、どのサイトも扱っていないものを検索対象にするためである。

【0028】前述の様にInitial Access Define ファイル6には、サイト1内のChild プロセス3が、常にデータ検索要求A₁を転送・発行するべきアドレス情報A₃が指定されているが、このInitial Access Define ファイル6は、以下の様なBNFで表される記述形式を持つ。

【0029】

【数3】

```

InitialAccessDefinefile ::= { <identifier> ' | ' } *;
<identifier> ::= <ip_string> | <dns_string> | <other>;
<ip_string> ::= { <number> 3 } { ' . ' <number> 3 } 3;
<dns_string> ::= { <word> ' . ' } { <word> ' . ' } *;
<other> ::= - <word> *;

```

【0030】次に、Routing Daemon プロセス2は、サイトデータ検索要求A₁を転送・発行するべきサイトのアドレス情報A₃、A₄にChild プロセス3を該当数分だけ立ち上げる。起動の際、各々のChild サイトプロセス3はデータ検索要求A₁を転送・発行するべきアドレス情報A₃、A₄、並びにデータ検索要求A₁の結果、転送されて来た内に保持されているデータ5'を、一時的に保持・記録するQuery Result Temp ファイル8の名称情報を引数A₅として受け取る。任意のChild プロセス3がサイト1'に該データ検索要求A₁と等価なデータ検索要求A₈を転送・発行する際は、サイト1'内のRouting Daemon プロセス2'が受理する。その際、データ検索要求A₈にはサイト1のアドレス情報が付加され、データ検索要求A₈内部のHop数が1つ追加される。

【0031】次に、Routing Daemon プロセス2は、自サイト内で管理しているデータ5からデータ検索要求A₁に応じたデータのサブセットA₆を受け取り、一時的に保持・記録する為、のQuery Result Temp ファイル8に該当データA₇として書き込む。Query Result Temp ファイル8は、以下のBNFで表される記述形式を持ち、プロセス3ごとに1つ作成される。

【0032】

【数4】

```

QueryResultTempFile ::= <header> { <URLdescription> * };
<header> ::= <querytext> <amount> ;
<amount> ::= <number> 6;
<URLdescription> ::= <URL> <object_identifier> <description> ;
<URL> ::= 'http://' { <ip_string> | <dns_string> } '/' <string> ;
<object_identifier> ::= <string> ;
<description> ::= <string> ;

```

【0033】サイト1'内でもサイト1と同様に、Routing Daemon プロセス2'が同じ処理を行ない、サイト1'内で管理しているデータ5'からデータ検索要求A8に応じたデータのサブセットA6を引き出し、転送データA9としてサイト内のChild プロセス3に戻す。Child プロセス3は転送データA9を受理すると、引数A5で指定されたQuery Result Temp ファイル8に該当データA10として書き込む。この処理は、起動されたChild プロセス3数分だけ実施される。

【0034】次に、Routing Daemon プロセス2は全てのChild プロセス3が該当データA10を書き込んだことを確認すると、全てのQuery Result Temp ファイル8をマージし、そこから該当データA11を読み出す。その後、Routing Daemon プロセス2は当該データA11をデータ検索要求A1に応じた転送データA14として、データ検索要求元に転送する。以上で、一連のデータ検索要求A1は完了する。Routing Daemon プロセス2は、不要となったQuery Result Temp ファイル8を消去する際に、Cached ファイル メンテナンス9へ起動要求17を送付する。Cached ファイル メンテナンス9は起動すると、データ検索結果を格納したQuery Result Temp ファイル8から、Categorized Access Cached

ファイル7上の相対網羅度A15、相対新規度A16を更新するためのパラメータデータA12を引き出し、これを基に更新元データA13を作成し、Categorized Access Cached ファイル7上の現相対網羅度A15、現相対新規度A16を更新する。前述の様にCategorized Access Cached ファイル7とは、データ検索要求A1、A8を転送した結果、得られる一時的なデータ検索格納先であるQuery Result Temp ファイル8から、事前に定義した当該データのマイニング方式に応じて、各サイトの評価を行ない、そのアドレス情報を管理するものである。Categorized Access Cached ファイル7により、Routing Daemon プロセス2は、データ検索要求A1、A8を転送・発行する際に、検索条件として指定されるカテゴリに対し最も有効と思われるサイトのみに転送・発行先を絞り込むことが出来るので、全サイトへ単純にブロードキャスト転送を行なうよりは、遥かに効果的なルーティング転送・発行処理を実現することが出来る。Categorized Access Cached ファイル7は以下のBNFで表される記述形式を持つ。

【0035】

【数5】

```

CategorizedAccessCachedfile ::= { <registry description> ' | ' } * ;
<registry description> ::= <category identifier> <identifier> <metrix description> ;
<category identifier> ::= <string> ;
<identifier> ::= <ip_string> | <dns_string> | <other> ;
<ip_string> ::= { <number> 3 } { ' . ' <number> 3 } 3;
<dns_string> ::= { <word> ' . ' } { <word> ' . ' } * ;
<other> ::= <word> * ;
<metrix description> ::= <relative covering degree> <relative strange degree> <counter> ;
<relative covering degree> ::= <number> 6;
<relative strange degree> ::= <number> 6;
<counter> ::= <number> 6;

```

【0036】Categorized Access Cached ファイル7上で管理される相対網羅度A15、相対新規度A16は、以下に記されるものである。

【0037】1) 相対網羅度A15

Cached ファイル メンテナンス9は当該データのマイニング方式としてサイト1'を始めとする関連する複数のサイトがデータ検索要求A1、A8に相当する任意データ検索要求を受理した結果として戻す転送データA9、A14に対し、各サイトがどの程度、検索条件である指定カテゴリに該当するものを含んでいるかを計算

し、更新元データA13を求める。その後、その計算値である更新元データA13で、Categorized Access Cached ファイル7上で管理されるサイト、カテゴリ毎に記録している相対網羅度A15の値を更新する。相対網羅度A15の評価に関しては後述する。

【0038】2) 相対新規度A16

Cached ファイル メンテナンス9は、前述の当該データのマイニング方式として、サイト1'を始めとする関連する複数のサイトがデータ検索要求A1、A8に相当する任意データ検索要求を受理した結果として戻す転送

データ A 9. A 1 4 に対し、各サイトがどの程度、検索条件である指定カテゴリに該当するもので独自かつ固有なものを含んでいるかを計算し、更新元データ A 1 3 を求める。その後その計算値である更新元データ A 1 3 で、Categorized Access Cached ファイル 7 上で管理される毎に記録している相対新規度 A 1 6 の値を更新する。相対新規度 A 1 6 の評価式に関しては後述する。

【0039】次に、当該データのマイニング方式について説明する。図 3 は、該当データのマイニング方式のフローチャートである。ステップ S 1 では、初期処理として、計算上、扱うべきサイト数を制限する。これは転送データ A 9 を戻すサイトは基本的に不特定多数であることから配慮されている。これを「サイト上限値ステップ」と呼ぶ。

【0040】ステップ S 2 は Child プロセス 3 毎に作成される Query Result Temp ファイル 8 から、〈header〉記述を集める処理である。〈header〉記述内部には、データ件数を意味する〈amount〉が記されている。

【0041】ステップ S 3 では、〈amount〉記述中の値で大きい順にソートして〈header〉記述を並べ替え、ステップ S 1 で制限されたサイト数に相当する順位のもの迄の複数サイトを処理の対象として選択する。これを「該当選択ステップ」と呼ぶ。

【0042】ステップ S 4 では、選択した複数サイトに対応する Query Result Temp ファイル 8 中の〈URLdescription〉記述を全て取り出し、1 つの〈URLdescription〉記述を 1 レコードと見なして、単純に連結する。なお、〈URLdescription〉記述の内部には、検索結果の対象を意味する〈object identifier〉記述が記されており、各々が 1 つの検索対象と見なされる。

【0043】ステップ S 5 では、〈object identifier〉記述でソートし、同じ〈object identifier〉記述を持つ異なるサイトの数を求める。以上を異なる〈object identifier〉記述毎全てにわたり実施する（ステップ S 6）。これを「該当 object 選択ステップ」と呼ぶ。

【0044】ステップ S 7 では、ステップ S 3 で選択した全サイトに対して（ステップ S 1 6）、各々、サイト数を変数 X とし、自身を含めた X 台のサイトに含まれる〈object identifier〉記述の度数 m (X) に関するヒストグラムを作成する。このステップ S 6 では、ステップ S 3 で選択した全サイトにわたり実施する。これを「ヒストグラム作成ステップ」と呼ぶ。

【0045】ステップ S 8 では、S 7 で求めたヒストグラムを基に網羅度及び新規度を以下の様に計算する。これを「網羅度・新規度算出ステップ」と呼ぶ。ステップ S 8 も、ステップ 7 と同様に、ステップ S 3 で選択した全サイトに対して行なう（ステップ S 1 7）。

【0046】

【数 6】

$$\text{網羅度} = \sum x \{m(X) * X\}$$

$$\text{新規度} = \sum x \{m(X)/X\}$$

【0047】ステップ S 9 では、ステップ S 8 で求めた網羅度の小さい順に選択したサイトをソートして、並べ替える。その後、小さい順に相対網羅度を 1 位、2 位、3 位と付与する。これを「相対網羅度付与ステップ」と呼ぶ。

【0048】ステップ S 1 0 では、ステップ S 8 で求めた新規度の小さい順に選択したサイトをソートして、並べ替える。その後、小さい順に相対新規度を 1 位、2 位、3 位と付与する。これを「相対新規度付与ステップ」と呼ぶ。

【0049】ステップ S 1 1 では、ステップ S 9 並びにステップ S 1 0 で求めた相対網羅度 A 1 5、相対新規度 A 1 6 を更新値 A 1 3 とする。その後、Categorized Access Cached ファイル 7 上でサイト、カテゴリ毎に管理されている現在の相対網羅度 A 1 5、現在の相対新規度 A 1 6 を読み出す。これを「既存データ読みだしステップ」と呼ぶ。

【0050】ステップ S 1 3 では、以下の処理を行なう。ステップ S 1 1 で現在の相対網羅度 A 1 5、現在の相対新規度 A 1 6 が読み出せる場合、該当する各々と更新値 A 1 3 との相加平均を取り、新たな相対網羅度 A 1 5、相対新規度 A 1 6 を求める。その際、相加平均の対象数を表す〈Counter〉記述値に 1 を加える。その後、Categorized Access Cached ファイル 7 に新たに算出した相対網羅度 A 1 5、並びに新たに算出した相対新規度 A 1 6 並びに〈Counter〉記述を書き込み、更新する。

【0051】ステップ S 1 1 で現在の相対網羅度 A 1 5、現在の相対新規度 A 1 6 が読み出せない場合は、ステップ S 9 で求めた相対網羅度 A 1 5、ステップ S 9 で求めた相対新規度 A 1 6、並びに該〈Counter〉記述を 1 として、Categorized Access Cached ファイル 7 に書き込み、更新する（ステップ S 1 4）。これを「更新ステップ」と呼ぶ。

【0052】ステップ S 1 1 並びに S 1 3、S 1 4 は、ステップ S 3 で選択した全サイト数回、処理を行う（ステップ S 1 5）。その後、規定回数分の処理を終えたならば、該当データのマイニング方式全体の処理を終了する。

【0053】

【発明の効果】本発明によれば、従来のように、語彙検索のみを対象とはしていないため、数値表現を含んだデータの検索が可能であり、一般的なレガシーデータをもその対象とすることが可能であるという第 1 の効果を有する。

【0054】また、本発明では、データのマイニング方式を採用しているため、動的にルーティングを実施出来、その結果、新たなサイトが登録され、それが該当デ

ータのマイニング方式の評価に対して妥当なデータを戻す場合は、新たなルーティング対象として、これを取り込むことが可能であるという第2の効果を有する。

【0055】さらに、本発明は、データ検索要求の起動状況を管理・記録するInvoke テーブルが実装されているので、検索が2重になることはないという第3の効果も有する。

【図面の簡単な説明】

【図1】 本発明のデータ検索装置の一実施例の構成図。

【図2】 本発明におけるルーティング先を決定する際の手順を示すフローチャート。

【図3】 本発明における該当データのマイニング方式の手順を示すフローチャート。

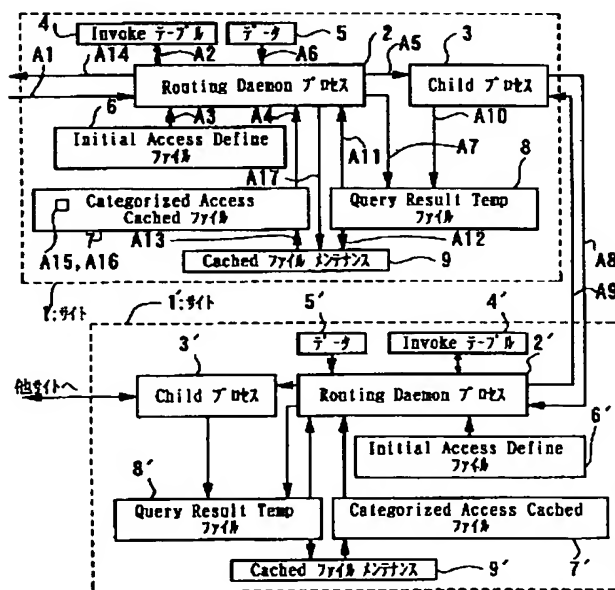
【図4】 従来方式の概要図。

【符号の説明】

1 データ転送装置（サイト）

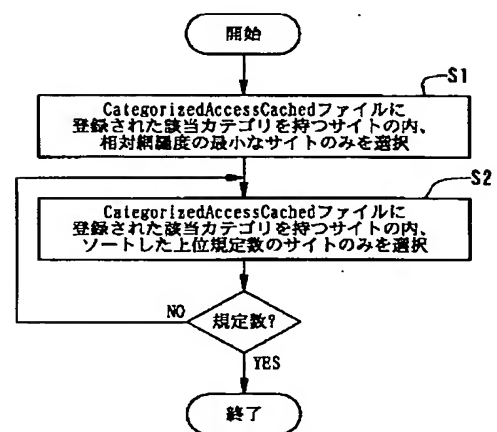
- 2 Routing Daemon プロセス
- 3 Child プロセス
- 4 Invoke テーブル
- 5 データ
- 6 Initial Access Define ファイル
- 7 Categorized Access Cached ファイル
- 8 Query Result Temp ファイル
- 9 Cached ファイル メンテナンス
- 11 Httpd プロセス
- 12 コンテントルータ
- 13 ルーティングデータベース
- 14 Refinement データベース
- 15 コンテンツラベル
- 16 サーチモジュール
- 17 WAIS プロセス
- 18 WAIS データベース

【図1】

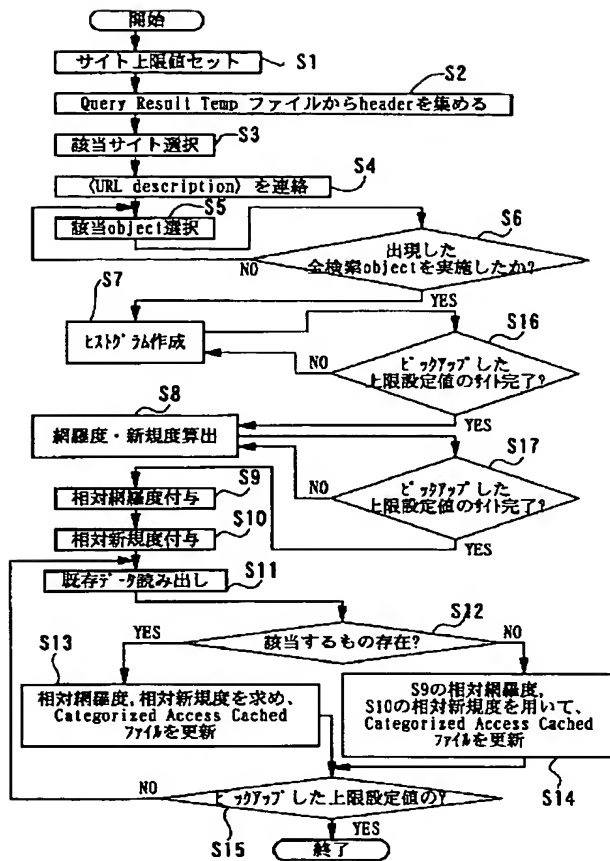


- | | |
|-------------|-------------|
| A1: データ検索要求 | A11: 該当データ |
| A2: 情報 | A12: パラメータ |
| A3: 子プロセス情報 | A13: 更新データ |
| A4: 子プロセス情報 | A14: 転送データ |
| A5: 引数 | A15: 相対網羅度値 |
| A6: サブセット | A16: 相対新規度値 |
| A7: 該当データ | A16: 起動要求 |
| A8: データ検索要求 | |
| A9: 転送データ | |
| A10: 該当データ | |

【図2】



【図3】



【図4】

